

COVARIANCE

$$COV(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

CORRELATION

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

CORRELATION

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

\bar{x} = mean of the values of the x-variable

y_i = values of the y-variable in a sample

\bar{y} = mean of the values of the y-variable

Range	Strength of association
0	No association
0 to ± 0.25	Negligible association
± 0.25 to ± 0.50	Weak association
± 0.50 to ± 0.75	Moderate association
± 0.75 to ± 1	Very strong association
± 1	Perfect association

OUTLIER

A DATA VALUE THAT IS NUMERICALLY DISTANT FROM A DATA SET



OUTLIER



A DATA VALUE THAT IS NUMERICALLY DISTANT FROM A DATA SET

**THE MEAN IS AFFECTED BY THE
PRESENCE OF OUTLIERS**

A DATA VALUE IS CONSIDERED TO BE AN OUTLIER IF..

DATA VALUE



$Q1 - 1.5(IQR)$

OR

DATA VALUE



$Q3 + 1.5(IQR)$

FIVE NUMBER SUMMARY

10 25 33 36 59

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

A DATA VALUE IS AN OUTLIER IF IT IS

LESS THAN

8.5

GREATER THAN

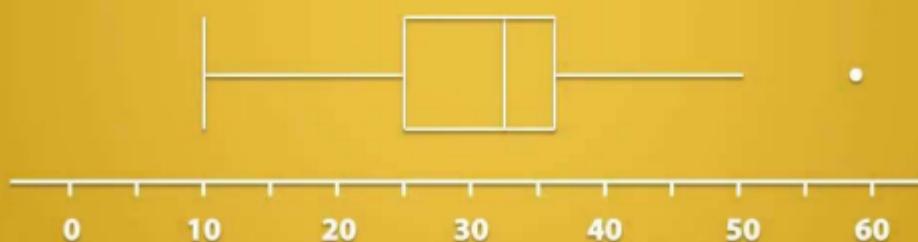
52.5

IQR = 11

FIVE NUMBER SUMMARY

10 25 33 36 59

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

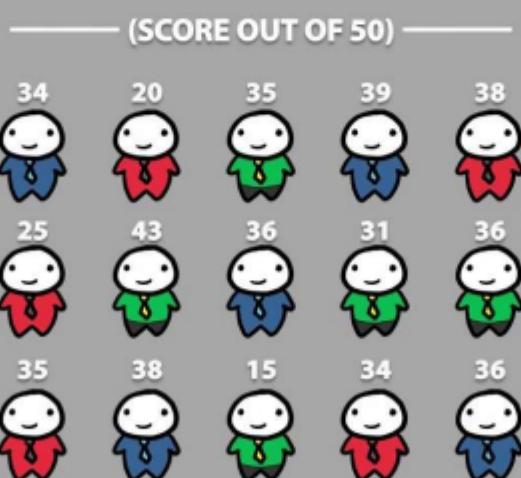


WHAT'S A **PERCENTILE?**

DESCRIBES THE PERCENTAGE OF
DATA VALUES THAT FALL AT OR
BELOW ANOTHER DATA VALUE



CLASS SIZE
 $n = 15$



50th percentile



50% OF THE DATA POINTS ARE AS SMALL OR SMALLER

Percentiles: Computational Procedure

- Organize the data into an ascending ordered array
- Calculate the p th percentile location:
$$i = \frac{P}{100}(n)$$
- Determine the percentile's location and its value.
- If i is a **whole number**, the percentile is the average of the values at the i and $(i+1)$ positions
- If i is **not a whole number**, the percentile is at the $(i+1)$ position in the ordered array

Percentiles: Example

- Raw Data: 14, 12, 19, 23, 5, 13, 28, 17
- Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28
- Location of 30th percentile:

$$i = \frac{30}{100}(8) = 2.4$$

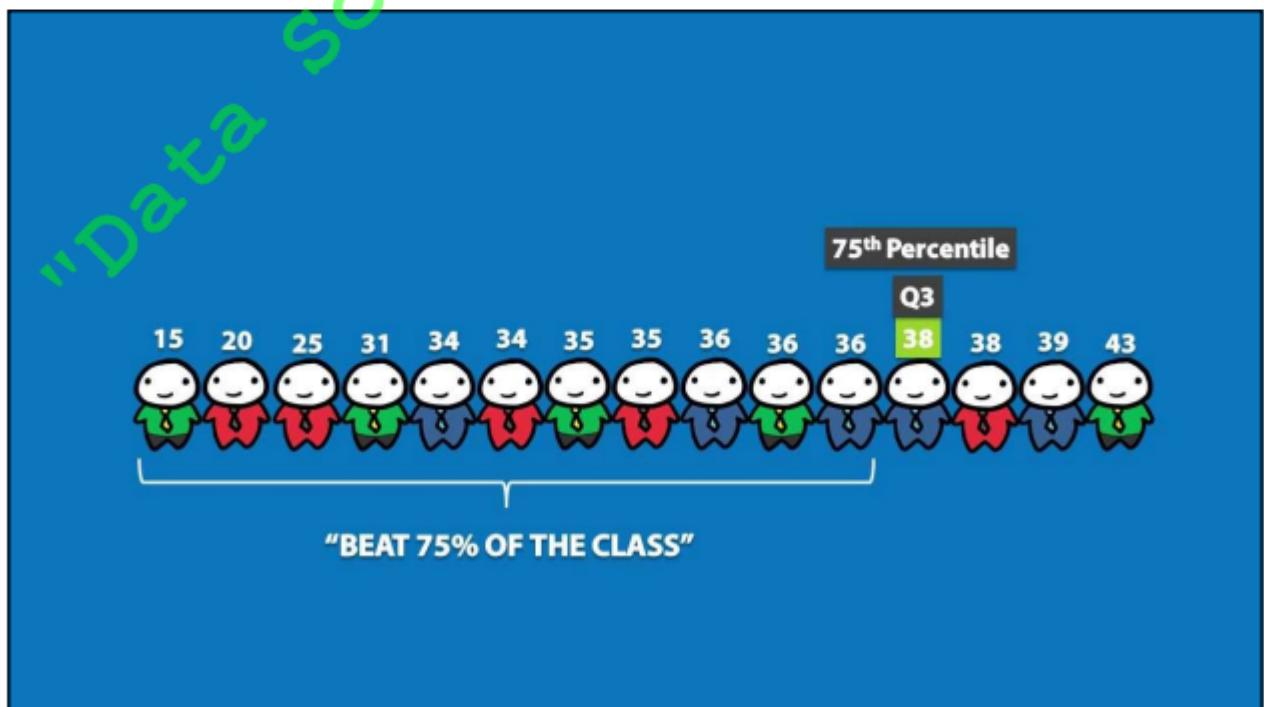
- The location index, i , is not a whole number; $i+1 = 2.4+1=3.4$; the whole number portion is 3; the 30th percentile is at the 3rd location of the array; the 30th percentile is 13.

Quartiles



Quartiles

- Measures of central tendency that divide a group of data into four subgroups
- Q_1 : 25% of the data set is below the first quartile
- Q_2 : 50% of the data set is below the second quartile
- Q_3 : 75% of the data set is below the third quartile
- Q_1 is equal to the 25th percentile
- Q_2 is located at 50th percentile and equals the median
- Q_3 is equal to the 75th percentile
- Quartile values are not necessarily members of the data set





Quartiles: Example

- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129

- $Q_1: i = \frac{25}{100}(8) = 2 \quad Q_1 = \frac{109 + 114}{2} = 111.5$

- $Q_2:$

$$i = \frac{50}{100}(8) = 4 \quad Q_2 = \frac{116 + 121}{2} = 118.5$$

- $Q_3:$

$$i = \frac{75}{100}(8) = 6 \quad Q_3 = \frac{122 + 125}{2} = 123.5$$

Interquartile Range

- Range of values between the first and third quartiles
- Range of the “middle half”
- Less influenced by extremes

$$\text{Interquartile Range} = Q_3 - Q_1$$

"Data Science & AI" by SRK

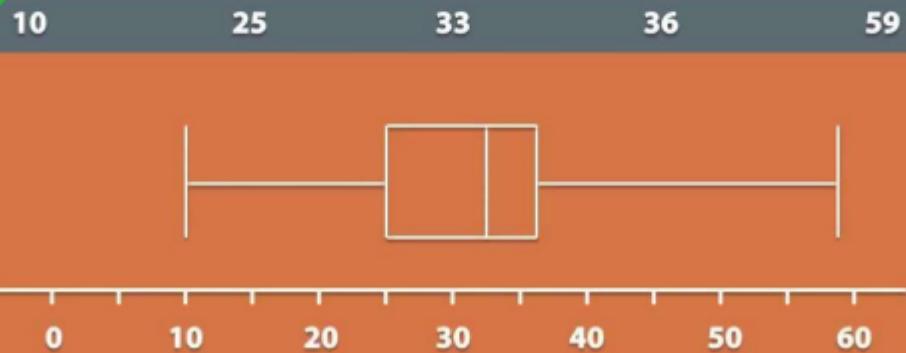
FIVE NUMBER SUMMARY
GIVES US A WAY TO DESCRIBE A DISTRIBUTION
USING ONLY **FIVE NUMBERS**

FIVE NUMBER SUMMARY

MINIMUM	1 ST QUARTILE	MEDIAN	3 RD QUARTILE	MAXIMUM
10	25	33	36	59

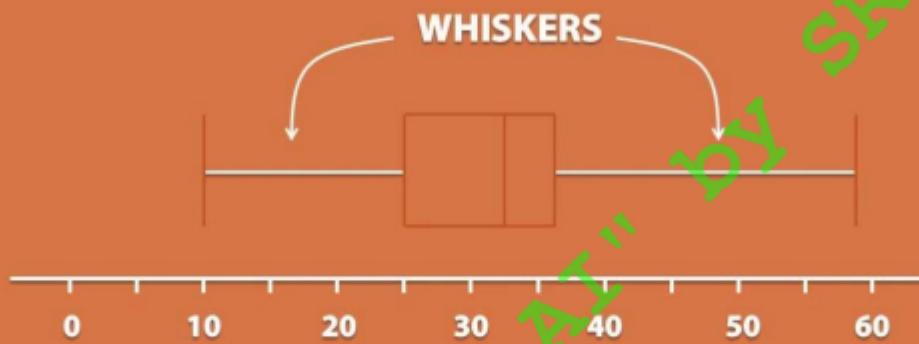
BOXPLOT

GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY



BOXPLOT

GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY



STRATEGIES FOR DETERMINING THE SKEWNESS FOR A BOXPLOT

UNEQUAL BOXES



SKEWED TO THE LEFT

EQUAL BOXES



SKEWED TO THE RIGHT

EQUAL BOXES WITH THE SAME WHISKER LENGTH



SYMMETRICAL

MEAS

C

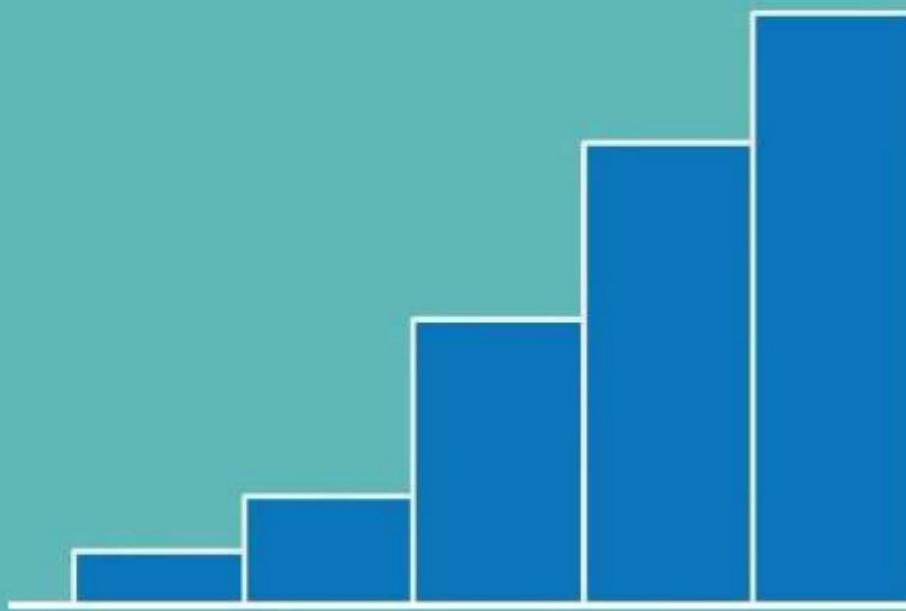
SHA

data science

data science

R

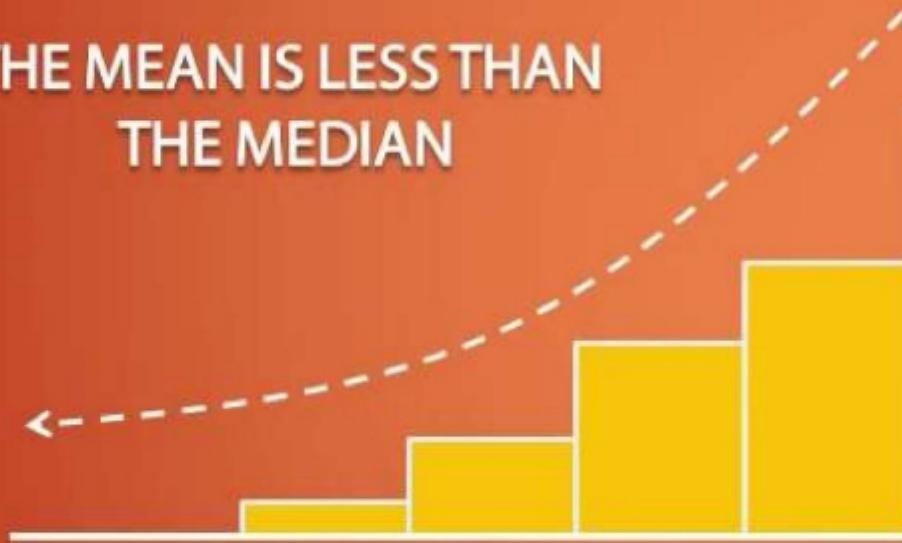
DISTRIBUTION:



Data Science

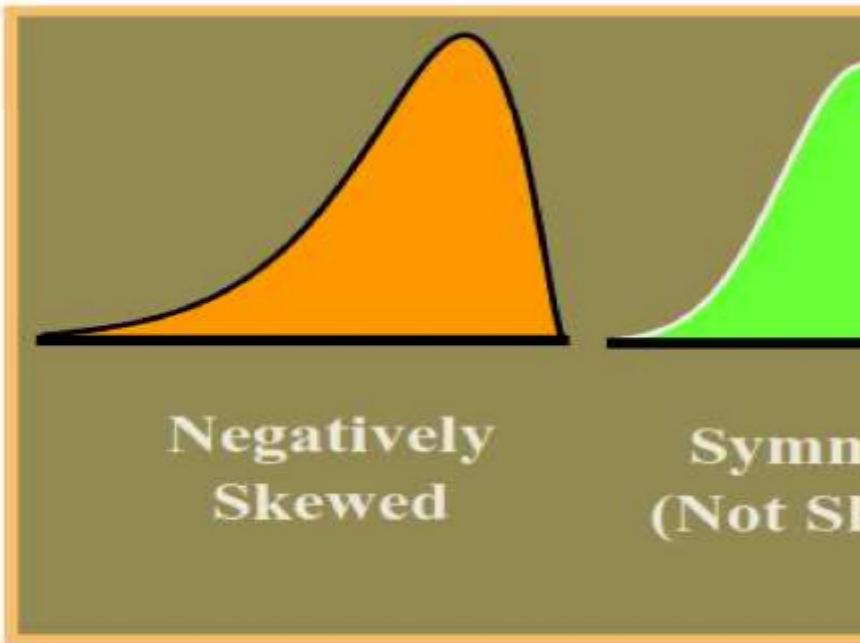
SKEWED TO THE LEFT

THE MEAN IS LESS THAN
THE MEDIAN



Data Science

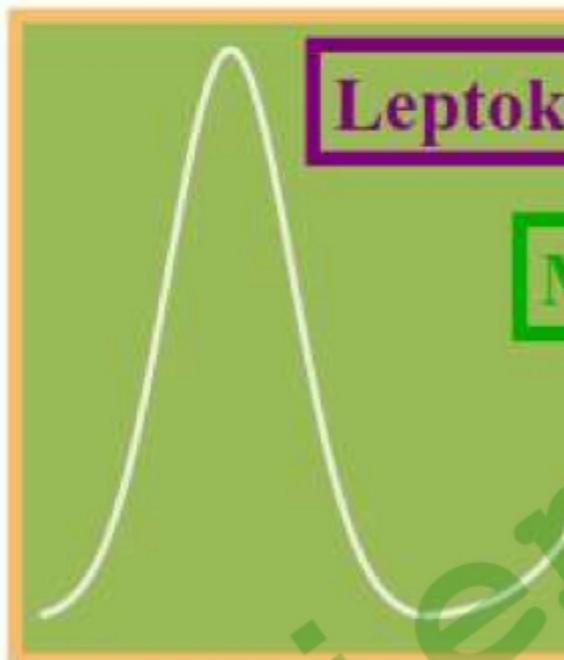
Skew



- If $S < 0$, the distribution is negatively skewed.
- If $S = 0$, the distribution is symmetric.
- If $S > 0$, the distribution is positively skewed.

Kurt

- Peakedness of a distribution
 - Leptokurtic: high and thin
 - Mesokurtic: normal in shape
 - Platykurtic: flat and spread out



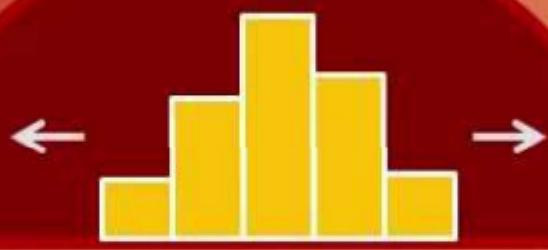
Data Science

MEAS

C

DISPERSION

*Acta Scientiarum
Parvorum*



MEASURES OF SPREAD

RANGE

STANDARD DEVIATION

Mean Absolute Deviation

- Average of the absolute deviation

X	$X - \mu$	$ X - \mu $
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

Population Standard Deviation

- Square root of the variance

X	$X - \mu$	$(X - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

Sample

- Average of the squared deviations

X	$X - \bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,825

STANDARD

HOW CLOSE THE DATA ARE TO THE MEAN

Data Science

STANDARD

HOW CLOSE THE DATA ARE TO THE MEAN



MEAN

HIGH

Data Science

Coefficient of Variation

$$\mu_1 = 29$$

$$\sigma_1 = 4.6$$

$$C.V_1 = \frac{\sigma_1}{\mu_1} (100)$$

$$= \frac{4.6}{29} (100)$$

$$= 15.86$$

MEAS

C

CENTRAL

data science

data science

R



154cm



MEASURES OF CENTRE

MODE = 154

REFERS TO THE DATA VALUE
THAT IS MOST FREQUENTLY
OBSERVED



MEASURES OF CENTRE

REFERS TO THE DATA VALUE
THAT IS POSITIONED IN THE

MEDIAN

MIDDLE OF AN ORDERED
DATA SET



MEASURES OF CENTRE

REFERS TO THE DATA VALUE
THAT IS POSITIONED IN THE

MEDIAN

MIDDLE OF AN ORDERED
DATA SET

Population Mean

$$\begin{aligned}\mu &= \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} \\ &= \frac{24 + 13 + 19 + 26 + 11}{5} \\ &= \frac{93}{5} \\ &= 18.6\end{aligned}$$

Weighted Average

- Sometimes we wish to average numbers according to their relative importance, or weight, to some other factor.
- The average you need is the **weighted average**.

Weighted Average

where x is a data value, w is the weight assigned to x , and n is the number of data values. The sum is the sum of all data values.

